DOCUMENT RESUME

ED 435 202 FL 026 056

AUTHOR Dunkel, Patricia A.
TITLE Considerations in Developing and Using Computer-Adaptive Tests To Assess Second Language Proficiency. ERIC Digest.
INSTITUTION ERIC Clearinghouse on Languages and Linguistics, Washington, DC.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO EDO-FL-99-07
PUB DATE 1999-10-00
NOTE 4p.
CONTRACT ED-99-CO-0008
AVAILABLE FROM ERIC/CLL, 4646 40th Street NW, Washington, DC 20016-1859. Tel: 202-362-0700; Web site: <http://www.cal.org/ericcll>.
PUB TYPE ERIC Publications (071) -- ERIC Digests in Full Text (073)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Elementary Secondary Education; Evaluation Methods; *Language Proficiency; *Language Tests; Second Language Instruction; Second Language Learning; *Test Items; *Test Reliability; *Test Validity; Testing
IDENTIFIERS ERIC Digests

ABSTRACT

Second language (L2) computer-adaptive testing (CAT) is a technologically advanced method of assessment in which the computer selects and presents test items to examinees according to the estimated level of the examinee's language ability. The basic notion of an adaptive test is to mimic automatically what a wise examiner would normally do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier. This approach stems from the realization that little is learned about an individual's ability if the questions posed are far too difficult or far too easy for that person. This digest discusses the following issues related to computer adaptive testing: (1) advantages of using computer-adaptive testing in second language assessment; (2) roots and challenges of computer-adaptive testing; (3) issues involving the basic principles of assessment in computer-adaptive testing; (4) CAT and the basic principles of test reliability; (5) CAT and the basic principles of test validity; and (6) CAT item pools for measuring the identified ability of the examinee population. (Contains 9 references.) (VWL)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 435 202

ERIC DIGEST

# Considerations in Developing and Using Computer-Adaptive Tests to Assess Second Language Proficiency

PATRICIA A. DUNKEL, GEORGIA STATE UNIVESITY

FL026056

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

ERIC DIGEST EDO-FL-99-07 OCTOBER 1999

# Considerations in Developing and Using Computer-Adaptive Tests to Assess Second Language Proficiency

PATRICIA A. DUNKEL, GEORGIA STATE UNIVERSITY

Today, powerful microcomputers are not only beginning to affect a redesign of the structure and content of school curricula and the entire process of instruction and learning, they are also having a decided impact on the types of tests created and used to assess that learning. In fact, computerized testing is increasingly being viewed as a practical alternative to paper-and-pencil testing (Kingsbury & Houser, 1993). Tests administered at computer terminals or on personal computers are known as computerized tests. Given the advantages of individual, time-independent language testing, computer-based testing will no doubt prove to be a positive development in assessment practice (Brown, 1997, p. 46).

## Computer-Adaptive Testing and Second Language Assessment

Second language (L2) computer-adaptive testing (CAT) is a technologically advanced method of assessment in which the computer selects and presents test items to examinees according to the estimated level of the examinee's language ability. The basic notion of an adaptive test is to mimic automatically what a wise examiner would normally do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier. This approach stems from the realization that we learn little about an individual's ability if we persist in asking questions that are far too difficult or far too easy for that person. We learn the most about an examinee's ability when we accurately direct our questions at the current level of the examinee's ability (Wainer, 1990, p. 10).

Thus, in a CAT, the first item is usually of a medium-difficulty level for the test population. An examinee who responds correctly will then receive a more difficult item. An examinee who misses the first item will be given an easier question. And so it goes, with the computer algorithm adjusting the selection of the items interactively to the successful or failed responses of the test taker.

## Advantages of Using Computer-Adaptive Testing in Second Language Assessment

In a CAT, each examinee takes a unique test that is tailored to his or her ability level. Avoided are questions that have low information value about the test taker's proficiency. The result of this approach is higher precision across a wider range of ability levels (Carlson, 1994, p. 218). In fact, CAT was developed to eliminate the time-consuming and inefficient (and traditional) test that presents easy questions to high-ability persons and excessively difficult questions to low-ability testees. Other advantages of CAT include the following:

- **Self-Pacing.** CAT allows test takers to work at their own pace. The speed of examinee responses could be used as additional information in assessing proficiency, if desired and warranted.
- **Challenge.** Test takers are challenged by test items at an appropriate level; they are not discouraged or annoyed by items that are far above or below their ability level.
- **Immediate Feedback.** The test can be scored immediately, providing instantaneous feedback for the examinees.
- **Improved Test Security.** The computer contains the entire item pool, rather than merely those specific items that will make up the examinee's test. As a result, it is more difficult to artificially boost one's scores by merely learning a few items or even types of items (Wainer, 1990). However, in order to achieve improved security, the item pool must be sufficiently large to ensure that test items do not reappear with a frequency sufficient to allow examinees to memorize them.
- **Multimedia Presentation.** Tests can include text, graphics, photographs, and even full-motion video clips, although multimedia CAT development is still in its infancy.

Individual test takers are not the only ones who can benefit from CAT. Green et al. (1995) point out that computerized testing could benefit a variety of agencies and groups, such as those described below:

- Educators considering the use of a published or in-house CAT to assess student achievement in large-enrollment second language classrooms or programs.
- Licensing boards needing to develop a CAT to help them identify candidates who meet specific performance standards for licensure. One such CAT is the Occupational English Test (OET) developed on behalf of the Australian Government (McNamara, 1991).
- Agencies preparing user guides for their computer-adaptive achievement tests, such as ETS's Graduate Record Examination.
- Departments of education wishing to develop a CAT version of statewide minimum competency tests.
- Departments of modern foreign languages wanting to create a proficiency CAT for entrance into or exit from required language courses, such as the Ohio State University's Multimedia Computer-Adaptive Test (MultiCAT) of French, German, and Spanish.

## Computer-Adaptive Testing: Roots and Challenges

In the 1960s and 1970s, the U.S. Department of Defense perceived the potential benefits of adaptive testing and supported extensive theoretical research in CAT and Item Response Theory (IRT), the family of psychometric models underlying computer-adaptive testing (Wainer, 1990, p. 10). IRT is based on probabilistic theory; that is, it calculates the probability of a given person getting a particular item right (Alderson, Clapham, & Wall, 1995). Examinees' scores and item total statistics are transformed into one scale so that they can be related to each other. If a person's ability is the same as the difficulty level of the item, that person has a 50-50 chance of getting that item right. If their ability is at a lower level than that of the item, that probability decreases. The relationship between the examinee's item performance and the abilities underlying item performance is described in an item characteristic curve (ICC). As the level of students' ability increases, so does the probability of a correct response (see Alderson, Clapham, & Wall, 1995, p. 90).

Early attempts to build adaptive tests by the U.S. Army, Navy, and Air Force were often less than successful, very expensive, and used large-scale computers. However, by the early 1980s, personal computers had acquired the power of the large-scale computers of earlier years, and the pioneering efforts of IRT theorists had perfected the psychometric model underlying CAT. In the late 1980s, CAT finally moved out of the realm of theory and supposition into the sphere of possibility and implementation with the advent of the College Board's CAT Graduate Record Examination and with the work of in-house researchers in foreign language education at the Defense Language Institute and at universities throughout the United States, Britain, the Netherlands, and other countries.

Today, with software development companies assisting test developers with their own institutional L2 CATs, computer-adaptive testing has finally become a viable alternative to conventional paper-and-pencil testing. Commercial CAT programs such as those offered by the Assessment Systems Corporation (St. Paul, Minnesota) and Computer-Adaptive Technologies (Chicago, Illinois) make it easier for developers



ERIC CLEARINGHOUSE ON LANGUAGES AND LINGUISTICS • CENTER FOR APPLIED LINGUISTICS • 4646 40TH ST NW • WASHINGTON DC 20016-1859 • (202) 362-0700

to create L2 CATs using software templates rather than having to start programming and development from scratch. It is anticipated that in the future, more and more commercial companies and academic institutions will be producing testing shells that can be used to create CATs for placement, achievement, and licensing purposes.

## Issues Involving the Basic Principles of Assessment in Computer-Adaptive Testing

A number of questions need to be addressed when considering the basic principles of assessment in computer-adaptive testing.

### Is the Computerized Testing System Appropriate for the Purpose of the Test?

L2 CAT developers need to clearly identify and specify the assessment purpose of their tests. This is important because CATs can be used for a wide variety of purposes, including the following:

- Identifying whether an individual has met the specific objectives of a basic language or literature course.
- Indicating an individual's level of achievement in a skill domain (e.g., listening comprehension or grammar knowledge).
- Identifying specific areas in which a student needs additional educational experiences (e.g., knowledge and use of specific grammatical points or recognition of specific idioms and vocabulary items).
- Diagnosing an individual's skill-area strengths (e.g., the ability to recognize main ideas presented in a spoken mini-lecture) and weaknesses (e.g., inability to recall specific details from a short conversation about an academic topic).
- Detecting whether candidates have met minimum course requirements as demonstrated in a mastery test.

In addition to clearly stating the purposes of the test, CAT developers must ensure that the CAT is able to measure the examinee's true proficiency level (Green et al., 1995). To achieve this goal, an L2 CAT must provide examinees with a sufficiently broad range of L2 content areas and skill tasks to ensure that their true proficiency is indeed measured in the test items taken. Because examinees may be of high or low proficiency levels, the CAT must be designed in such a way as to provide adequate assessment for the entire range of ability represented in the examinee population (Green et al., 1995, p. 2). This objective may most easily be accomplished by obtaining or designing a CAT that includes the entire range of ability in its item pool. For example, in the case of a general listening proficiency CAT, items in the pool must cover low to high listening ability levels. In addition, the items need to include a variety of listening tasks, such as comprehension of the main ideas of a conversation or mini-lecture, recognition and recall of details of a conversation, identification of specific words and phrases used in a passage, and so forth.

To achieve both objectives, the item selection algorithm must constrain the selection of items not just on the basis of the statistical parameter associated with the test item (such as the difficulty level), but it must also be able to present a variety of designated listening comprehension tasks to the examinees.

### Does the CAT Embody the Basic Principles of Reliability?

Reliability refers to the precision and consistency of scores derived from a test instrument. It is a function of general, situational, and individual factors (Cohen, 1994) that can be used to frame evaluative questions for the developers of the test. General factors influencing reliability include, for example, whether instructions for the examinees are clear and explicit, or whether the examinees are sufficiently familiar with the format of the CAT before taking it. Situational factors include those related to the testing environment, such as noise level or whether headphones are provided. Individual factors include transient factors, such as the physical and psychological health of the test takers, and stable factors, such as examinees' experience with similar tests.

### Does the CAT Embody the Basic Principles of Validity?

Validity refers to whether a test actually measures what it purports to measure. It relates to the appropriacy of the inferences made on the basis of the test scores. There are several aspects of validity: content, construct, criterion, concurrent, and predictive. CAT developers and users need to examine issues related to each of these types of validity.

### Do the Examinees Have an Opportunity to Become Familiar with the Computer, the CAT System, and the Structure, Organization, and Content Domains of the CAT?

Examinees should be given the time and opportunity to become thoroughly familiar with both the computer and the testing system. For first-time computer users, there should be an orientation to the functioning of the computer (e.g., using a mouse, calling for questions, answering questions, adjusting the audio volume, scrolling, etc.). An orientation to the structure and types of items they will encounter during the CAT should also be required for all examinees. The practice items should be equivalent in structure and content to those contained in the item bank.

### Is the Item Pool of an Appropriate Quality to Support the Test Purpose(s) and to Measure the Identified Ability of the Examinee Population?

The depth and breadth of the item pool from which individual items are drawn strongly affects the validity and utility of the resulting CAT scores. Because of this, in addition to ensuring that the items tap the variety of specific tasks and content areas pertinent to the identified purpose of the CAT, the developers and users of the scores need to be able to specify exactly what the items in the bank assess. For instance, in an academic listening proficiency CAT, the designers could specify that all examinees must demonstrate comprehension of the main ideas of a mini-lecture and comprehension of the details of a short dialog. They may also wish to set other specific skills for certain ability levels. For instance, advanced listeners should be able to understand the implied meaning of utterances.

## Conclusion

Computer-adaptive testing shows promise in becoming a regular component of standardized foreign language assessment in the coming century, particularly for licensing and certification purposes. Many benefits accrue to examinees and administrators alike when using CAT. However, to reap these benefits, numerous checks and balances need to be put into place so that the potential pitfalls in the development and uninformed use of CAT are avoided. Developers and users alike need to understand fully what a CAT is and how it operates. They also need to be aware of what the underlying psychometric model used in their CAT posits in terms of the unidimensional or multidimensional IRT model selected. They need to understand what the selected IRT model means in terms of the dimensionality of the content and tasks associated with the items. They need to be familiar with how the IRT statistical parameters of the test items are estimated after their trialing. Above all, they must know what is necessary to implement a valid and reliable CAT.

## References


Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* New York: Cambridge University Press.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1,* 44-59. http://polyglot.cal.msu.edu/llt/vol1num1/brown/default.html

Carslon, R. (1994). Computer-adaptive testing: A shift in the evaluation paradigm. *Journal of Educational Technology Systems, 22,* 213-224.

Cohen, A. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle & Heinle.

Green, B., Kingsbury, G., Lloyd, B., Mills, C., Plake, B., Skaggs, G., Stevenson, J., Zara, T., & Schwartz, J. (1995). *Guidelines for computerized-adaptive test development and use in education.* Washington, DC: American Council on Education, Credit by Examination Program.

Kingsbury, G., & Houser, R. (1993). Assessing the utility of item response models: Computer adaptive testing. Educational Measurement: *Issues and Practice, 12,* 21-27.

McNamara, T. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8,* 139-159.

Wainer, H. (1990). *Computer adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum.



The full version of this article appeared in *Language Learning & Technology* (Vol. 2, No. 2, January 1999, pp. 77-93) http://polyglot.cal.msu.edu/llt/vol2num2/





This digest was prepared with funding from the U.S. Dept. of Education, Office of Educational Research and Improvement, National Library of Education, under contract no. ED-99-CO-0008. The opinions expressed do not necessarily reflect the positions or policies of ED, OERI, or NLE.